Information Technology Laboratory

Information Access Division (IAD)

RT-02 Metadata Scoring

A. Martin

for the NIST Gang

May 7, 2002

Rich Transcription ’02 Workshop

What is MetaData?

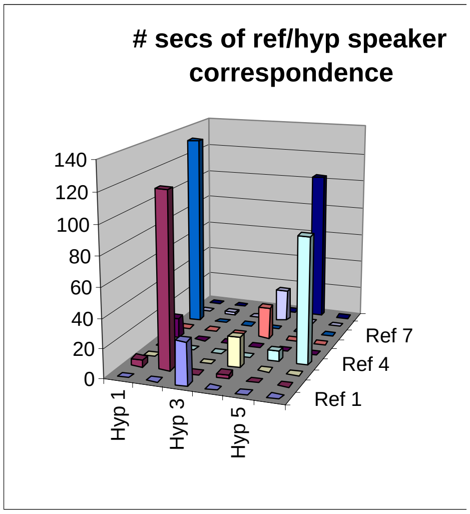
- A working definition
  - Metadata is non-orthographic information which can be extracted from audio (and/or video) signals (J. Garofolo)
- Possible examples
  - Sentences, paragraphs, punctuation
  - Channel, environment
  - Dialog info
  - Topic
  - …
- NIST ran an experiment to define usable types
  - Details tomorrow
- For RT-02 used
  - Speaker identities: task is to cluster speech data by individual (unknown) speakers

Scoring

- Same problem investigated (and scored) previously in speaker recognition evaluations
  - speaker segmentation task
- Intervals of no speech or overlapping speech ignored for scoring
- Also ignored “collar” intervals of 0.25 seconds around endpoints of speech intervals as annotated
- Scoring software seeks best one-to-one mapping of actual (reference) speakers and each system’s hypothesized speakers

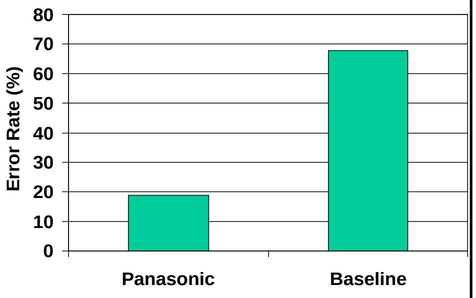
Scoring (cont’d)

- Example (from broadcast news)
  - 9 reference speakers, 6 hypothesized speakers
  - Form matrix of overlapping speech durations
- Pick one-to-one mapping with maximum sum
- Error rate is accumulated total “leftover” speech durations divided by total speech duration



Results: Broadcast News

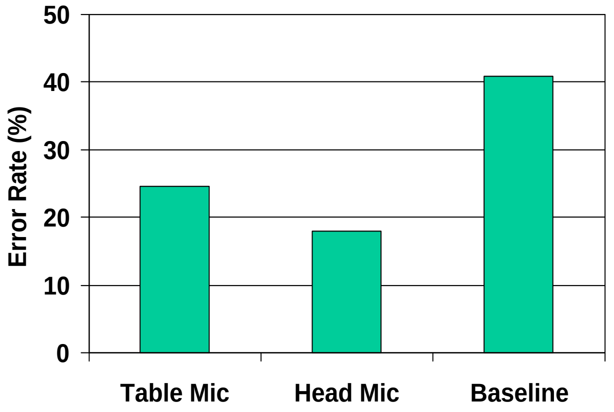
- Panasonic was the only participant
  - – Total speech: 3268.95 sec.
  - – Correctly matched 2652.99 sec.
  - – Incorrectly matched 615.96 sec.
  - – Error rate: 18.8%
- How much information is actually gained?
  - Compare with *Baseline* (knowledge-free) error rate which associates all speech with a single speaker
  - Panasonic's error rate was reduced by more than two thirds from the baseline



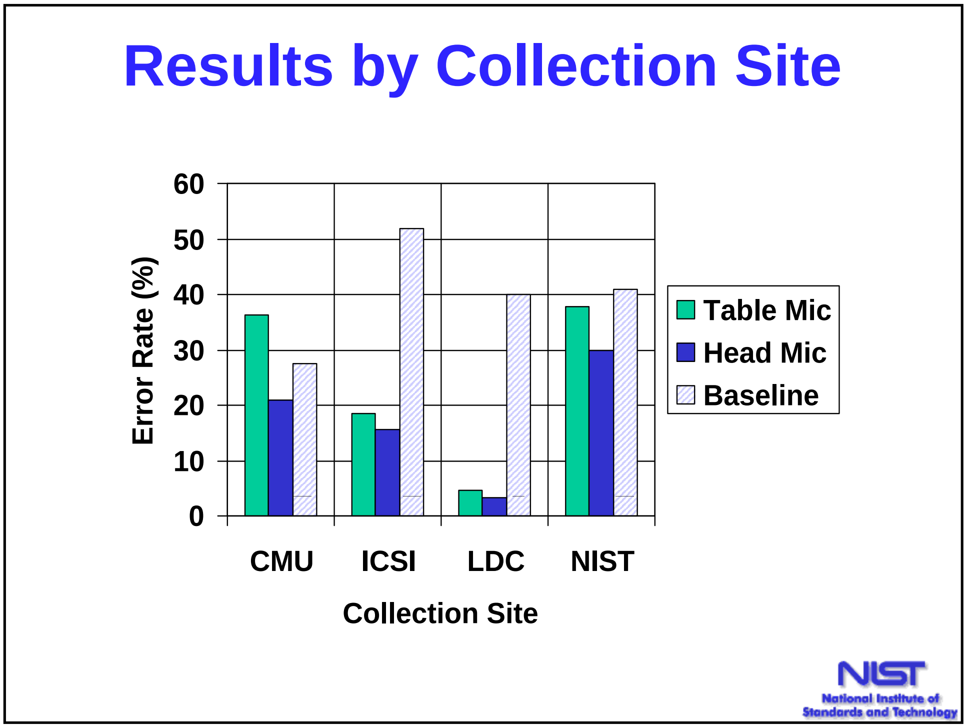
National Institute of Standards and Technology

Results: Meetings

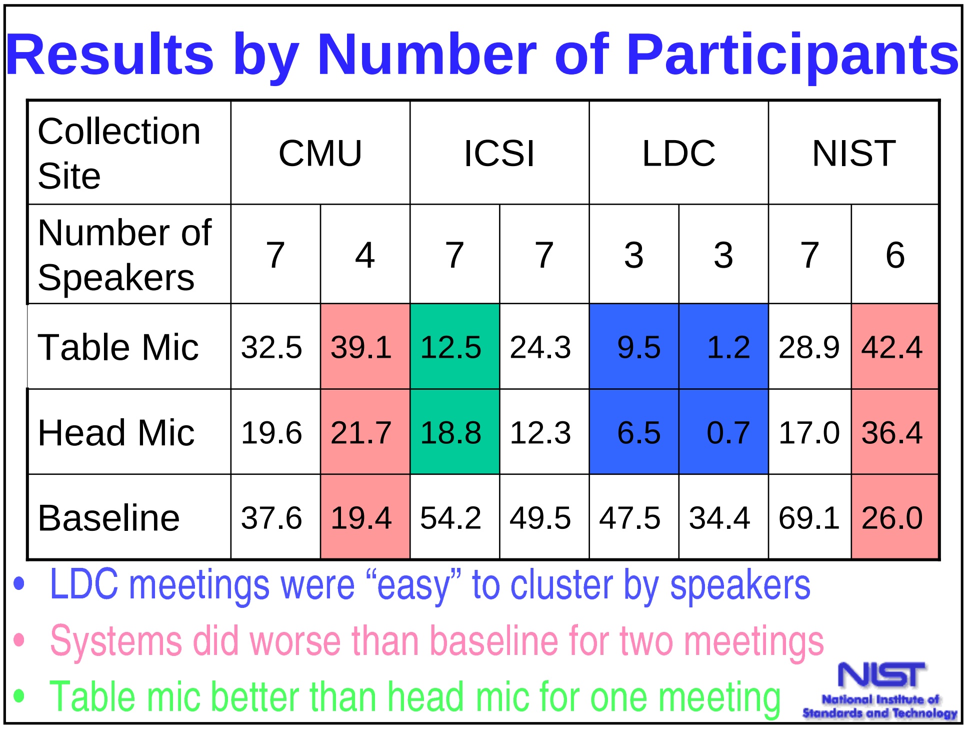
- MITLL1: center table mic (contrastive mic)
- MITLL2: head mounted mic (control mic)



National Institute of Standards and Technology



Results by Collection Site



Results by Number of Participants

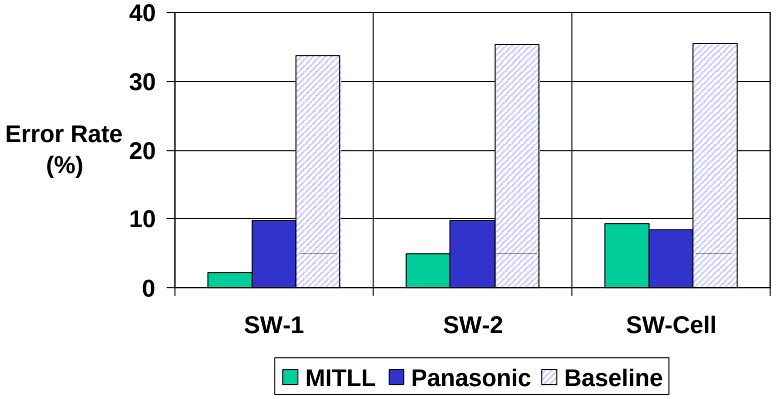
Collection Site	CMU		ICSI		LDC		NIST	
Number of Speakers	7	4	7	7	3	3	7	6
Table Mic	32.5	39.1	12.5	24.3	9.5	1.2	28.9	42.4
Head Mic	19.6	21.7	18.8	12.3	6.5	0.7	17.0	36.4
Baseline	37.6	19.4	54.2	49.5	47.5	34.4	69.1	26.0

- LDC meetings were “easy” to cluster by speakers
- Systems did worse than baseline for two meetings
- Table mic better than head mic for one meeting

NIST
National Institute of Standards and Technology

Results: Telephone Conversations (using summed channel data)

- MITLL and Panasonic each participated



Issues

- Overlapping speech
  - Ignored this year, but real systems must deal with this
- Speaker segmentation from text
  - May use ASR transcript in addition to the audio signal (as has been done in the speaker detection task)
- Scoring
  - Should detecting speech/non-speech be included in scoring, or is speech activity detection a non-problem?
  - Speaker boundary detection could be separately scored (as in TDT story boundary detection)
  - Could reformulate error score in terms of speaker substitutions, misses, and false alarms, as done this year in the speaker recognition evaluation, which would allow different weights for different error types